U.S. DEPARTMENT OF ENERGY

Human Terrain Mapping (HTM) Data Review Process/Standard Operating Procedure

For Use by DOE Institutional Review Boards and Researchers in Reviewing Datasets for HTM Projects

May 2012

This Data Review Process was reviewed and concurred on by:

John C. Ordaz, NNSA HSP Program Manager (NA-SH-40) Date: 5/24/2012
DOE Management Team

Elizabeth P. White, DOE HSP Program Manager (SC-23.2) Date: 5/24/2012
DOE Management Team

This HTM Data Review Process was reviewed and approved by:

Sharlene Weatherwax, Ph.D., Associate Director for Biological and Environmental Research (SC-23) and DOE Institutional Official Date: 5/24/2012

HTM Data Review Process

Human Terrain Mapping (HTM) data often contains personally identifiable information (PII). DOE policy states HTM research activities are managed as human subjects research and only de-identified data (defined below) may be used. The purpose of de-identifying the data is to add another layer of separation between the DOE/DOE lab's principal investigator (PI) and the military's ultimate use of that data. Also, de-identifying the data protects (to a certain extent) the dataset, should there be an unintentional release.

The recognized DOE site IRB is the only entity authorized to determine whether the HTM data received by the PI after project initiation meets the DOE criteria for de-identification. If the DOE site does not manage or operate its IRB, then the Central DOE IRB shall be the responsible IRB.

> DOE defines ***de-identified data*** as "a data set that has no, or limited, identifiers and for which a person with current knowledge of generally accepted scientific principles determines that the risk that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient, to identify an individual who is a subject of the information, has been reduced to the extent practicable." A graded approach must be used in balancing de-identification of the datasets and the usability of the dataset to accomplish the needed research.

The HTM Data Review process has been developed to help the PI and the IRBs assure that datasets received by the investigator have been de-identified to the extent practicable while still allowing the PI to complete his/her work. The IRBs will provide oversight for the management and handling of HTM datasets, and will follow-up with the PI periodically.[1]

Core to the HTM Data Review process is a checklist intended to be used by PIs and IRBs to communicate the risk(s) posed by a research dataset. This checklist is the formal document used by the PI to communicate to the IRB areas of concerns, to describe the protective measures implemented, and to document due diligence in assuring that the data used is de-

[1]While subject to a site or Central DOE IRB review, DOE laboratory HTM research activities typically will be considered exempt human subjects research. The IRB will review the statement of work and approve the dataset(s) to be used, but won't typically conduct another formal review unless the PI is considering modifications to the scope of work or dataset(s). The IRB will communicate with the PI periodically regarding status and scope of the project.

identified to the extent practicable. The checklist, signed by the PI, will be submitted to the IRB. The IRB will provide a formal response to the PI indicating what changes must be made, if any, prior to start of the project. This checklist is applicable for HTM research projects of all sizes and any level of security classification.

Before proceeding to complete the HTM Data Review Checklist below, the PI should review the following steps and ensure they have been addressed.

Start of Human Terrain Mapping (HTM) Project

After the contract is awarded by the sponsor (note: sponsor refers to the institution funding the project, which, as an example, may be another Federal agency, DOE Program Office/Laboratory, University, etc.) for a project that was determined to be HTM by DOE Headquarters and approved by both DOE Headquarters and the applicable DOE, the following process must be followed:

A. Verification

1. **Approved Statement of Work** – The PI must ensure that the DOE Site Office, IRB, and appropriate DOE/NNSA HQ Human Subjects Protection (HSP) Program Manager have a copy of the final approved Statement of Work for the HTM project.

2. **Appropriately De-Identified Dataset** - If the sponsor committed to submit a de-identified dataset, the PI must review the dataset, once received, to verify that it has been de-identified. The PI should then complete the attached checklist, in preparation for the consultation with the DOE Site or Central IRB. ***Note:*** *The designated DOE IRB is the only entity authorized to determine if the dataset has been properly de-identified to meet DOE HTM policy.*

3. Inadequately De-Identified Dataset - If the PI believes the dataset has not been appropriately de-identified by the sponsor, s/he should contact the designated IRB and the appropriate DOE/NNSA HQ HSP Program Manager, as well as the DOE Site Office, before starting the project and distributing or using the dataset.

B. Sponsor Requested DOE De-Identification of Dataset - If the sponsor requested that the DOE lab de-identify the dataset(s) to be used for the project, it should have been done by another organization within that DOE lab, with a certain degree of organizational separation, with limited communication between the two organizations, and under a separate contractual task. Specifically:

1. **DOE Coordination -** The appropriate DOE/NNSA HSP Program Manager will coordinate with the PI, the DOE Site Office, and the Site or Central IRB, as needed, and potentially also the sponsor, to ensure that an appropriate team is selected to de-identify the datasets.

2. **Independent De-Identification Team** - An appropriate de-identification team is one whose members will not be engaged in the conduct of the research and whose organizations do not have a first level common manager.

3. **De-Identification Task** - The de-identification/re-identification services must be provided under a separate contractual task order with the sponsor and it must not be managed by the research PI.

4. **Communications** - The identified dataset will not be shared with the PI/researchers who will perform work on the sponsor's task. The team de-identifying/re-identifying the data will sign a non-disclosure agreement documenting this degree of independence. However, limited communications, if needed, may take place between the team de-identifying and/or re-identifying the sponsor's data and the PI. Such communications will be solely for the purpose of clarifying the data the PI will need in order to successfully conduct the project.

5. **Transmittal of De-Identified Dataset** - Once de-identified, the team responsible for this task will send the dataset(s) back to the sponsor for review/approval prior to project initiation. The sponsor will then send the approved dataset to the PI to begin work.

Upon completion of the independent de-identification and receipt of the dataset(s), the PI should follow steps A.1. through A.3., above, as applicable.

The HTM Data Review Checklist should be completed by the PI in preparation for meeting the IRB.

Note: This is a "living document" intended to be updated and revised as feedback on its utility and effectiveness is received. Comments and suggestions are encouraged and will be reviewed to inform ongoing revisions to this document.

DOE CHECKLIST

REQUIREMENTS FOR PROTECTING PERSONALLY IDENTIFIABLE INFORMATION (PII)

1. Keep PII confidential;
2. Release PII, where required, only under a procedure approved by the responsible IRB(s) and DOE;
3. Use PII only for purposes of this program;
4. Handle and mark documents containing PII as "containing PII or PHI;
5. Establish reasonable administrative, technical, and physical safeguards to prevent unauthorized use or disclosure of PII;
6. Make no further use or disclosure of the PII except when approved by the responsible IRB(s) and DOE, where applicable, and then only under the following circumstances: (a) in an emergency affecting the health or safety of any individual; (b) for use in another research project under these same conditions and with DOE written authorization; (c) for disclosure to a person authorized by the DOE program office for the purpose of an audit related to the project; (d) when required by law; or (e) with the consent of the participant.
7. Protect PII data stored on removable media (CD, DVD, USB Flash Drives, etc.) using encryption products that are Federal Information Processing Standards (FIPS) 140-2 certified;
8. Use passwords to protect PII used in conjunction with FIPS 140-2 certified encryption that meet the current DOE password requirements cited in DOE Guide 205.3-1;
9. Send removable media containing PII, as required, by express overnight service with signature and tracking capability, and shipping hard copy documents double wrapped;
10. Encrypt data files containing PII that are being sent by e-mail with FIPS 140-2 certified encryption products;
11. Send passwords that are used to encrypt data files containing PII separately from the encrypted data file, i.e. separate e-mail, telephone call, separate letter;
12. Use FIPS 140-2 certified encryption methods for websites established for the submission of information that includes PII;
13. Use two-factor authentication for logon access control for remote access to systems and databases that contain PII. (Two-factor authentication is contained in the National Institute of Standards and Technology (NIST) Special Publication 800-63 Version 1.0.2 found at: http://csrc.nist.gov/publication/nistpubs/800-63/SP800-63V 1 0 2.pdf);
14. Report the loss or suspected loss of PII immediately upon discovery to: 1) the DOE funding office Program Manager; and 2) the applicable IRBs (as designated by the DOE Program Manager). If the DOE Program Manager is unreachable, immediately notify the DOE-CIRC (1-866-941-2472, www.doecirc.energy.gov).

Mark as **Official Use Only** after checklist is completed

Human Terrain Mapping Data Review Checklist

This checklist is intended to be completed by the PI and reviewed by the IRB for the purpose of assuring that the dataset(s) to be used have been sufficiently de-identified, to the extent practicable, for the PI to begin work. This checklist should be completed before bringing it to the IRB for approval. However, please feel free to contact your IRB Program Manager (Administrator) for help at any point along the way.

Assumptions:

1) The dataset(s) under review is/are for a research project that was determined by DOE to be HTM.
2) Once the IRB approves the dataset(s), a Data Security Agreement will be jointly completed by the PI and the IRB.

Data Description	
Name of project	

Name of dataset	

Size of dataset (e.g., GB)	

Format (e.g., database, image, xml, spreadsheet, etc.)	

Content (e.g., text, graphics, biomedical, email, domestic vs. international, etc.)	

Source (e.g., sponsor, original, purchased, etc.)	

Sensitivity level of dataset (e.g., OUO, classified, etc.) *If the dataset is classified, consult IRB before completing remainder of checklist.*	

Identifiable Information	
After receiving the dataset from the sponsor, describe what data has been de-identified or removed and how it was de-identified. *Please note that the process to determine if the dataset has identifiable information begins upon receipt of the data from the sponsor. The sponsor should have been made aware of the DOE Policy of only accepting and working with sufficiently de-identified datasets*	

Has the sponsor left any identifiers in the dataset? If yes, - **describe the identifiable data.** - **If this identifiable data is needed, can it be de-identified? If not, please explain.** - **If this identifiable data is not needed, can it be deleted or de-identified? If not, please explain.** *Please note that DOE requires that only de-identified datasets be used for HTM projects, so this data must be de-identified to the extent practicable.*	

Describe any plans to merge this data with other datasets and indicate if the combined dataset will have identifiable data, and whether such data is publicly available or not. If so, describe how the aggregate dataset will be managed.	

Data Protection Plan		
Access Control	**Discuss your data access protection plan** (e.g., who will have access to this data?)	
	Discuss your data distribution plan (especially as it pertains to persons external to your project)	
	What computer systems and networks can access this data?	
	Will this data be published? If so, where, and do you have the sponsor's approval?	

Storage	**Discuss your data storage plan** (e.g., how will this data be stored?)	

Encryption	**Will the data be encrypted when not in use?**	

Post-project	**What will happen to this data when the project ends?**	

Overall Assessment	
Are there any concerns in adhering to "DOE's requirements for protecting personally identifiable information (PII) (attached) , even for this de-identified (or almost fully de-identified dataset)?	

Describe your confidence level in the de-identification applied to this dataset. (e.g., minimal quasi-identifiers present, dataset size small enough for thorough review, etc.)	

Describe your confidence level in the protection planned for this dataset. (Consider both internal and external risks such as probability of data exposure, secure data access, etc.)	

Describe other known risks that were not sufficiently addressed above.	

Principal Investigator's Name Signature Date

Appendix A

Glossary of Terms

De-identified Data - A data set that has no, or limited, identifiers and for which a person with current knowledge of generally accepted scientific principles determines that the risk that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient, to identify an individual who is a subject of the information, has been reduced to the extent practicable. A graded approach must be used in balancing de-identification of the datasets and the usability of the dataset to accomplish the needed research.

Extraneous Data – Information in the dataset not relevant/needed for the specific research to be undertaken that may contain Personally Identifiable Information.

Human Subjects Research (HSR) – Any systematic investigation (including research development, testing, and evaluation) involving intervention or interaction with individuals or using their personally identifiable information or materials, designed to develop or contribute to generalizable knowledge. In addition to traditional biomedical and clinical studies, such research includes but is not limited to studies that –

(1) Use humans to examine devices, products or materials with the express purpose of investigating human-machine interfaces or evaluating environmental alterations when humans are the subjects being tested;
(2) Use personally identifiable bodily materials such as cells, blood, tissues, urine, or hair, even if the materials were collected previously for a purpose other than the current research;
(3) Collect and use personally identifiable information such as genetic information or medical and exposure records, even if the information was collected previously for a purpose other than the current research;
(4) Collect personally identifiable or non-identifiable data, surveys, or questionnaires through direct intervention or interaction with individuals; and
(5) Search for generalizable knowledge about categories or classes of subjects (e.g., linking job conditions of worker populations to hazardous or adverse health outcomes).

Human Terrain Mapping (HTM) - Research and data gathering activities primarily conducted for military or intelligence purposes to understand the "human terrain"— the social, ethnographic, cultural, and political elements of the people among whom the U.S. Armed Forces are operating and/or in countries prone to political instability. This work includes observations, questionnaires, and interviews of groups of individuals, as well as modeling and analysis of collected data, and may become the basis for U.S. military actions in such locations. In addition to Human Terrain Mapping (HTM), such activities are often referred to as human social culture behavior (HSCB) and human terrain systems (HTS) studies. It is DOE policy that HTM activities will be managed as HSR.

HTM Data - Data collected or used as part of HTM efforts, as described above, as well as any auxiliary data on the same group(s) of individuals.

Identifier – See Appendix B.

Institutional Review Board (IRB) - A committee or board established by an institution that performs initial and continuing reviews of research involving human subjects, and is registered with the Office for Human Research Protections (OHRP) and designated on a Federal Wide Assurance (FWA).

Internet research is any human subjects research conducted using the Internet. On the internet are two types of information: *publicly available* and *for authorized use only*.

Publicly Available**:** Information is publicly available when it is lawfully made available to the general public from: (1) Federal, state, or local government records; (ii) Widely distributed media, including information that has been published or broadcast for public consumption, is accessible online to the public, or is available to the public by subscription or purchase; or (iii) Disclosures to the general public that are required to be made by federal, state, or local law. Publicly available does not mean "without restriction" (see note below).

For Authorized Use Only: Information that is restricted to authorized users and governed by specific data protection rules.

Note: All internet research, regardless of information type, must comply with the appropriate DOE directives, such as level of security/classification and protection of personally identifiable information (PII). Only information obtained with due authorizations and that complies with applicable requirements will be approved by DOE IRBs/HSPP. The applicable DOE site IRB is the only entity authorized to approve the information to be used. If the DOE site does not manage or operate its IRB, then the Central DOE IRB shall be the responsible IRB.

Merged Data - Data from two or more datasets that has been combined into a single new data set.

Personally Identifiable Information - Any information collected or maintained about an individual, including but not limited to, education, financial transactions, medical history and criminal or employment history, and information that can be used to distinguish or trace an individual's identity, such as his/her name, Social Security number, date and place of birth, mother's maiden name, biometric data, and any other personal information that is linked or linkable to a specific individual. Refer to DOE O 206.1, *Department of Energy Privacy Program*.

Appendix B

Identifiers, Quasi-Identifiers and Data Security Requirements

Introduction

The purpose of this appendix is to provide a reference for IRB members and investigators on:

1) Identifiers: data that can uniquely identify individuals
2) Quasi-identifiers: data that does not explicitly identify individuals but when used in combination with other data can do so
3) Potential data security requirements: Requirements the IRB can ask to be enforced on data.

Identifiers

1. HIPAA "Safe Harbor" Fields

As of February 2, 2011, the Health Insurance Portability and Accountability Act of 1996 (HIPAA), as amended. Privacy and Security Rules indicates the following data be removed or de-identified before sharing medical data on individuals. Note that for HTM projects, some of these identifiers may not apply and are listed here as a reference.

Data Field	Notes
Names	
Geographic subdivisions	Any geographic subdivision less than a state. *
Dates (except year)	Includes birth date, admission date, discharge date, date of death, and ages > 89. **
Telephone /Fax numbers	
Email	
Social Security Numbers	
Medical Record numbers	
Health plan beneficiary numbers	

Account numbers	
Certificate/license numbers	
Vehicle identifiers (including license plate numbers)	
Device identifiers and serial numbers	
URLs	
IP Addresses	
Biometric identifiers (including finger and voice prints)	
Full face photographic images	

- *The initial three digits of the zipcode are allowed in certain cases. See 45 CFR §164.514 for more details.
- ** Ages above 89 can be aggregated into a >90 category.

Quasi-Identifiers

The following table contains quasi-identifiers sets that have been used to re-identify data sets. Each set has a reference that provides more detail. Re-identification based on quasi-identifier sets is dependent upon many variables (such as the existence and availability of an auxiliary data set) and thus this list should be taken primarily as a starting point for further discussion.

Quasi-Identifier sets	Note	Reference
Zip code, birth date, sex	References: Indicate that >50% of U.S. individuals have a unique combination of these fields.	Sweeney, 2000
Movie title, rating, date of movie ratings	Re-identification of some Netflix users based on these fields, using data from the public movie rating site IMDB.com	Narayanan, 2008
International Statistical Classification of Diseases and Related Health Problems (ICD-9)	Set of diagnosis codes for a patient can possibly be unique.	Loukides, 2010

Potential Data Security Requirements

Some data security requirements that the IRB may impose on a project as a function of the data properties (e.g., level of de-identification, size of data set) and project properties (e.g., utility of certain attributes) include:

- Data use only on systems disconnected from the network;
- When not in use, the dataset must be encrypted according to lab approved encryption program and/or on a storage device not physically connected to a machine; and
- Dataset must be destroyed or turned over to the sponsor at the end of the project.

References

L. Sweeney. Uniqueness of Simple Demographics in the U.S. Population. Technical Report LIDAP-WP4, Laboratory for International Data Privacy, 2000.

A. Narayanan and V. Shmatikov. Robust De-Anonymization of Large Sparse Datasets. In Proceedings of the 29th IEEE Symposium on Security and Privacy, 2008.

G. Loukides, J. C. Denny, and B. Malin. The Disclosure of Diagnosis Codes Can Breach Research Participants Privacy. Journal of the American Medical Informatics Association, 17:3.

Health Insurance Portability and Accountability Act of 1996 (HIPAA), as amended.

Mark as **Official Use Only** after HTM Data Security Agreement is completed

Appendix C

HTM Data Security Agreement

This HTM Data Security Agreement is entered on, _____,201_ between

_____ _____ ("IRB") and
IRB Chair (please print) *DOE Laboratory/Site*

_____ _____ ("PI").
Principal Investigator (please print) *Organization*

This Agreement establishes the terms and conditions under which the PI will protect the HTM data

_____ for the _____
HTM dataset (s) name *HTM research project name*

as approved for use by the IRB (see attached HTM Data Review Checklist). If a Data Security Agreement exists for the use of this HTM data between a Sponsor and the PI, this Agreement must complement and not contradict those terms and conditions.

Use of this data, in whole or in part, for other HTM and/or human subjects research projects will be subject to prior approval by the IRB.

The terms and conditions of this Agreement are developed jointly between the PI and IRB during the discussion about the HTM Data Review Checklist. This Agreement can be changed only by a written modification of the agreement by the party signatories (or their replacements) to this Agreement or by the parties adopting a new agreement in place of this Agreement.

The PI will be designated as custodian of this HTM data and will be responsible for complying with all conditions of use and for establishment and maintenance of security arrangements as specified in this Agreement to prevent unauthorized use and disclosure of this data.

Access to the data will be limited to the minimum number of individuals who need access to this data. Describe below how access will be limited and/or controlled.

The appropriate administrative, technical, and physical safeguards will be established to prevent unauthorized use or access to this data. Describe below what safeguards will be used to store, distribute, transmit, publish, etc. the data.

Upon project completion, the data will be archived/transferred/destroyed/etc. as described below, and the IRB notified once these steps have been completed.

Derivative data created from this data will be managed in the same manner as the original data, unless otherwise approved by the IRB.

The PI will inform the project team (to include team members, subcontractors, collaborators, and any other parties with access to the HTM dataset) of this HTM Data Security Agreement and will have them countersign below to document their awareness and expected compliance to these data security terms and conditions.

The PI agrees to notify the IRB immediately if the Agreement or any provision of this Agreement has been breached. Such notification will include the identity of such individuals and the nature of the breach.

The PI agrees to notify his/her management and the IRB if the project data has been lost, stolen, or compromised in any way. The response is expected to be in alignment with security incident reporting and the sponsor of the project will be notified as appropriate.

This Agreement may be terminated by either party at any time for any reason upon ten (10) days written notice. Upon such notice, the project data will be handled per the project completion plan described above.

This Agreement expires upon written notification by the PI to the IRB that the project has been completed and data properly disposed and/or secured (as described above). This Agreement will be transferred to and signed by a new PI after any change of custodianship.

_____	_____
PI Signature	*IRB Chair Signature*
_____	_____
Printed Name	Printed Name
_____	_____
Date	Date